Home Assignment
Course: Applied Statistics and Big Data - Module A1
Instructor: Dimitris Fouskakis
19/03/2018

1. Consider the Credit dataset that can be found in `blackboard`. The dataset contain the following variables:

Variable Name	Description
`Income`	Income in thousands of dollars
`Limit`	Credit limit in dollars
`Rating`	Credit rating
`Cards`	Number of credit cards
`Age`	Age in years
`Education`	Years of education
`Gender`	Gender (Male or Female)
`Student`	Student (Yes or No)
`Married`	Married (Yes or No)
`Ethnicity`	Ethnicity (African American, Asian, Caucasian)
`Balance`	Average credit card debt in dollars

   i. Load the data into R. Remove the first column, with the name `id`.

2. Consider the data of Exercise 1. Using R perform the following tasks.

   i. Fit a simple linear regression model using `Balance` as the response variable and `Limit` as the explanatory variable. Construct the plot of the least squares fitted line.

   ii. Check the assumptions of the above model.

   iii. Give the interpretation of the estimated coefficients of your linear model and perform statistical inference about them. Give an estimate of the standard deviation of the errors and comment on its value.

   iv. Fit a simple linear regression model using `Balance` as the response variable and `Ethnicity` as the explanatory variable. Give the interpretation of the estimated coefficients of your linear model.

   v. Fit a multiple linear regression model using `Balance` as the response variable and all the remaining variables as explanatory.

vi. Give the interpretation of the estimated coefficients of your multiple regression model and perform statistical inference about them. Give an estimate of the standard deviation of the errors and comment on its value.

vii. Based on the results of your last model, produce a 95% confidence interval for the expected `Balance` of an individual with income 30000$ with 3000$ credit limit, with credit rating 300, with 1 credit card, who is 25 years of age, female, with 7 years of education, a student, not married and Asian.

3. Consider the data of Exercise 1 and create a new categorical variable `Balance.cat` which takes the value 0 (low) when Balance is less than 454$ and the value 1 (high) in all other cases. Using R perform the following tasks.

   i. Fit a simple logistic regression model with `Balance.cat` as the response variable and `Limit` as the explanatory variable.

   ii. Produce the summary of the logit model you have fitted and interpret the estimated coefficients in terms of odds and odds ratios.

   iii. Give a 95% CI for the odds ratio corresponding to the above models coefficient of `Limit`. Give its interpretation.

   iv. Fit a multiple logistic regression model with `Balance.cat` as the response variable and `Limit` and `Ethnicity` as the explanatory variables.

   v. Produce the summary of the logit model you have fitted and interpret the estimated coefficients (in terms of odds ratios) of `Limit` and of the dummy variables of `Ethnicity`.

   vi. Give 95% CI's for the odds ratios corresponding to the last models coefficient of `Limit` and of the dummy variables of `Ethnicity` and interpret them appropriately.

   vii. Based on the results of your last model, estimate the probability of an African American individual having a low `Balance` when his/her credit limit is 4000$.

Instructions

i. **Assignment submission deadline**: 26 March, 2018 at 13:00 (Italian Time). Please send me your paper at fouskakis@math.ntua.gr. Please note that no assignment will be acceptable after this date and time.

ii. Your paper should be on a pdf format. This file should be named using the following format: SURNAME-NAME.pdf (replace with your details). The file should start with a cover page in where you will include your details (title of the assignment, your name, your surname, your email and your student number).

iii. In the pdf file you should try to present the solutions of the exercises in a compact way and explaining the interpretations of your findings as simple as possible. Also you should include the R code.

iv. The paper must be typed on a computer (no scanned) and its maximum length would be 20 pages. You can use any word processor you wish but you have to send me the pdf file at the end. All questions are compulsory.

v. It is important that the coursework reflects your knowledge rather than it being simply an accumulation of information. The assignment should be well structured and easy to read.